

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

SPATIAL AND SPATIOTEMPORAL POINT PROCESS MODELLING IN EPIDEMIOLOGY

A thesis presented in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

IN

STATISTICS

at Massey University, Manawatu

New Zealand

TILMAN MARCUS DAVIES

December, 2011

Abstract

Geographical epidemiology encapsulates those problems that aim to understand space and/or space time trends in the disease(s) of interest, a goal clearly important in both public health and economic contexts. The increasing availability of point-location, coordinate data for studies in geographical epidemiology calls for a greater scrutiny of statistical point process theory with respect to these applications. Though the statistical analysis of planar point patterns is now theoretically well-established, there remain many aspects which warrant further research, especially from a practical perspective. The need becomes even greater when we consider the relative youth of methods for the analysis of spatiotemporal observations, where non-trivial variation and even dependence throughout the space-time continuum can exist.

This work aims to address these issues by careful review, theoretical refinement, empirical testing, and real-world analyses of certain statistical tools used in point process problems in geographical epidemiology. We scrutinise the kernel-smoothed density-ratio estimator of the so called *relative risk function*, a particularly flexible approach given the anticipated spatial heterogeneity of the observations over a given geographical region. This discussion introduces the adaptive (i.e. variable bandwidth) risk function, as well as novel asymptotic methods for computation of tolerance contours designed to identify sub-regions of statistically significant fluctuations in risk.

More sophisticated statistical methodology is warranted in certain situations, where it may be assumed that both ‘global’ heterogeneity and ‘local’ correlation drives the space and/or space-time disease dispersion. A comprehensive review of the stochastic *log-Gaussian Cox process*, in both purely spatial and spatiotemporal contexts, is conducted. A suite of novel numerical experiments investigate the performance of convenient, yet *ad hoc*, minimum contrast parameter estimation techniques for the dependence structure of the latent Gaussian process.

The computer code arising from the review and refinement of the above methodologies was instrumental in the release of two separate software packages. These are available in the R environment, and also showcased here. A number of additional collaborations with applied researchers around the world serve to further highlight the contributions made throughout the course of this research project and the importance of sound statistical methods in geographical epidemiology.

Acknowledgements

To my amazing family, both across the Tasman and up in Germany, for their unwavering and continuous support;

To my wonderful cat, Sigma, for always being happy to see me and for maintaining a consistent supply of fluff and semi-deceased birds;

To the statistics academics across Europe for their interest in my work and generous hospitality during my trip in early 2011;

To my external collaborators of various peer-reviewed works that took an interest in the contents of this thesis while still in production; this academic experience at such an early stage of my career was indeed invaluable, and of course extremely rewarding. Special thanks are reserved for M. Stevenson, EpiCentre, Massey University (New Zealand), B. Taylor, B. Rowlingson and P. Diggle, Division of Medicine, Lancaster University (U.K.), and Z. Zhang, Department of Epidemiology, Fudan University (People's Republic of China);

To my local co-supervisors, Associate Professor Geoff Jones and Doctor Jonathan C. Marshall, a massive thanks for never failing to provide that extra support and additional points of view absolutely invaluable to understanding and implementing novel and/or complex material;

And most importantly to my primary supervisor, Professor Martin L. Hazelton, without whom this work quite simply would not exist. Responsible not just for this project's supervision, but for inspiring me in undergraduate years to further pursue statistics in the first place, he ensured that the positive times far outweighed the negative. True intelligence is not just knowledge, but the ability to communicate, and teach, said knowledge. Martin is, without a doubt, the most intelligent academic I've ever met.

Contents

List of Figures	ix
List of Tables	xiv
List of Publications	xv
List of Acronyms	xvi
1 Introduction	1
1.1 Data in geographical epidemiology	1
1.1.1 Illustrative examples	1
1.1.2 Preliminary discussion	4
1.2 Point processes	6
1.2.1 Complete spatial randomness	6
1.2.2 The intensity function	8
1.2.3 Second-order characteristics	11
1.2.4 Higher-order properties	17
1.3 Thesis aims and structure	17
2 Spatial Relative Risk	20
2.1 Bivariate kernel density estimation	21
2.1.1 Estimator	22
2.1.2 Theoretical properties	22
2.2 The relative risk function	25
2.2.1 Definition	25
2.2.2 Properties	26
2.3 Practical considerations	28
2.3.1 Bandwidths	28
2.3.2 Edge-correction	33
2.3.3 Kernel function	34
2.3.4 Trimming	36
2.4 Asymptotic tolerance contours	37
2.5 Simulation study	38
2.6 Examples	49

2.6.1	Chorley-Ribble cancer data	49
2.6.2	Primary biliary cirrhosis in northeastern England	52
2.7	Closing remarks	54
3	Spatial and Spatiotemporal Modelling with the Log-Gaussian Cox Process	57
3.1	Definitions and properties	58
3.1.1	Spatial	58
3.1.2	Spatiotemporal	60
3.1.3	The LGCP in epidemiology: deterministic components	62
3.1.4	An alternative model	65
3.2	Minimum contrast parameter estimation	66
3.2.1	Correlation structures	66
3.2.2	Concept of minimum contrast	68
3.2.3	Practical issues	71
3.2.4	Numerical likelihood-based alternatives	73
3.3	Computational considerations	74
3.3.1	General simulation: the fast-Fourier transform	74
3.3.2	Conditional simulation: the Metropolis-adjusted Langevin algorithm	80
3.4	Real-world examples	86
3.4.1	Humberside leukaemia and lymphoma	86
3.4.2	AEGISS gastrointestinal infections	91
3.5	Simulation studies	103
3.5.1	Study 1: K or g ?	103
3.5.2	Study 2: Deterministic bandwidth impact on spatial parameter estimation	107
3.5.3	Study 3: Spatiotemporal exceedance probability sensitivity	115
3.6	Potential theoretical refinements	126
3.6.1	Improved ‘ratio estimators’ of second-order characteristics	127
3.6.2	Multivariate LGCP modelling	128
3.6.3	Nonseparable spatiotemporal correlation structures	129
3.7	Summary and concluding remarks	130
4	Software Releases	134
4.1	R package sparr	134
4.1.1	Aims and scope	135
4.1.2	Example walkthrough and code snippets	136
4.1.3	Commentary regarding fast and efficient use	145
4.1.4	Limitations and future extensions	146
4.2	R package lgcp	147
4.2.1	Aims and scope	147
4.2.2	Example walkthrough and code snippets	149
4.2.3	Some interesting notes on the FFT	163
4.2.4	Limitations and future extensions	166

5 Collaborative Applications and Case Studies	169
5.1 Schistosomiasis in Guichi, China	169
5.1.1 Background	170
5.1.2 Data and methods	170
5.1.3 Summary of results	172
5.1.4 Conclusions	175
5.2 Avian influenza – global trends	175
5.2.1 Background	176
5.2.2 Data, methods, results	176
5.2.3 Concluding remarks	178
5.3 Foot-and-mouth disease model verification, Ireland	180
5.3.1 Project summary	180
5.3.2 Study design specifics	181
5.3.3 Analysis and results	183
5.3.4 Discussion	185
5.4 Spatiotemporal mapping of Japanese earthquakes, 1926–2005: a preliminary exploration	187
5.4.1 Data and analysis proposals	188
5.4.2 Model specification, parameter estimation	190
5.4.3 Concurrent prediction and future-forecasting	190
5.4.4 Closing comments, critiques	194
6 Discussion	198
6.1 Summary and accomplishments	199
6.1.1 Chapter 1	199
6.1.2 Chapter 2	199
6.1.3 Chapter 3	200
6.1.4 Chapter 4	201
6.1.5 Chapter 5	201
6.2 Future research, thesis closure	202
Appendices	206
A Package documentation: sparr	206
A.1 ‘Welcome’ help file: <code>sparr-package</code>	206
A.2 Conversion to pixel image class <code>im: as.im.bivden, as.im.rrs</code>	208
A.3 Bivariate (2D) kernel density estimation: <code>bivariate.density</code>	210
A.4 Single-density LSCV bandwidth: <code>LSCV.density</code>	215
A.5 Density-ratio LSCV bandwidth: <code>LSCV.risk</code>	217
A.6 Bivariate (2D) Gaussian kernel: <code>KBivN</code>	219
A.7 Bivariate (2D) quartic (biweight) kernel: <code>KBivQ</code>	220
A.8 Single-density normal scale rule bandwidth: <code>NS</code>	221
A.9 Terrell’s single-density oversmoothing bandwidth: <code>OS</code>	223
A.10 Primary biliary cirrhosis (PBC) data: <code>PBC</code>	224

A.11 Plotting density and risk function estimates: <code>plot.bivden</code> , <code>plot.rrs</code>	225
A.12 Relative risk function estimation: <code>risk</code>	227
A.13 Print summary of estimated density: <code>summary.bivden</code>	231
A.14 Print summary of estimated risk function: <code>summary.rrs</code>	231
A.15 Asymptotic and numerical tolerance contours (p -value surfaces): <code>tolerance</code> .	232
A.16 Internal functions: <code>sparr-internal</code>	235
B Package documentation (excerpt): <code>lgcp</code>	237
B.1 Space-time data set object (generic): <code>stppp</code>	237
B.2 Global spatial variation object (generic): <code>spatialAtRisk</code>	238
B.3 Global temporal variation object (generic): <code>temporalAtRisk</code>	239
B.4 Interactive kernel estimation of global spatial trend: <code>lambdaEst</code>	241
B.5 Nonparametric estimation of global temporal trend: <code>muEst</code>	242
B.6 Time-averaged inhomogeneous PCF: <code>ginhomAverage</code>	243
B.7 Time-averaged inhomogeneous K -function: <code>KinhomAverage</code>	244
B.8 Interactive spatial minimum contrast estimation: <code>spatialparsEst</code>	245
B.9 Interactive temporal contrast estimation: <code>thetaEst</code>	246
B.10 Conditional simulation of a space-time LGCP via the MALA: <code>lgcpPredict</code> .	247
B.11 Plotting posterior means from a MALA run: <code>plot.lgcpPredict</code>	249
Bibliography	251

List of Figures

1.1	Snow’s recorded locations of cholera deaths (\cdot) and water pumps (\times) during the 1854 outbreak in London.	2
1.2	Standardised mortality ratio (SMR) of respiratory cancer by county in Ohio, USA, for the year 1985.	3
1.3	Spatial locations of Burkitt’s lymphoma occurrences in the Western Nile district of Uganda, Africa, for the years 1961, 1968 and 1975.	4
1.4	Three hypothetical examples of 2D point patterns.	7
1.5	Estimates of Ripley’s K -function for the three point patterns A , B and C	14
1.6	Estimates of the PCF based on a homogeneous underlying intensity for the three point patterns A , B and C	16
2.1	Humberside case-control observations (left), along with the $LSCV_\rho$ function values for a sequence of bandwidths (center), and the same $LSCV_\rho$ function values on the log-scale (right).	31
2.2	Filled contour plots of the synthetic control densities and log-risk functions for Problems 1 to 5.	41
2.3	Filled contour plots of the synthetic control densities and log-risk functions for Problems 6 to 10.	42
2.4	Problem 1 log-scaled ISE measure for sample sizes $n_1 = 100$, $n_2 = 100$ (white) and $n_1 = 500$, $n_2 = 1000$ (grey).	43
2.5	Problem 2 log-scaled ISE (left) and WISE measures (right) for sample sizes $n_1 = 100$, $n_2 = 100$ (white) and $n_1 = 500$, $n_2 = 1000$ (grey).	44
2.6	Problem 3 log-scaled ISE (left) and WISE measures (right) for sample sizes $n_1 = 100$, $n_2 = 100$ (white) and $n_1 = 500$, $n_2 = 1000$ (grey).	44
2.7	Problem 4 log-scaled ISE (left) and WISE measures (right) for sample sizes $n_1 = 100$, $n_2 = 100$ (white) and $n_1 = 500$, $n_2 = 1000$ (grey).	45
2.8	Problem 5 log-scaled ISE (left) and WISE (right) measures for sample sizes $n_1 = 100$, $n_2 = 100$ (white) and $n_1 = 500$, $n_2 = 1000$ (grey).	45
2.9	Problem 6 log-scaled ISE (left) and WISE (right) measures for sample sizes $n_1 = 100$, $n_2 = 100$ (white) and $n_1 = 500$, $n_2 = 1000$ (grey).	46
2.10	Problem 7 log-scaled ISE (left) and WISE (right) measures for sample sizes $n_1 = 100$, $n_2 = 100$ (white) and $n_1 = 500$, $n_2 = 1000$ (grey).	46
2.11	Problem 8 log-scaled ISE (left) and WISE (right) measures for sample sizes $n_1 = 100$, $n_2 = 100$ (white) and $n_1 = 500$, $n_2 = 1000$ (grey).	47

2.12	Problem 9 log-scaled ISE (left) and WISE (right) measures for sample sizes $n_1 = 100, n_2 = 100$ (white) and $n_1 = 500, n_2 = 1000$ (grey).	47
2.13	Problem 10 log-scaled ISE (left) and WISE (right) measures for sample sizes $n_1 = 100, n_2 = 100$ (white) and $n_1 = 500, n_2 = 1000$ (grey).	48
2.14	Laryngeal cancer cases (left) and lung cancer controls (right) for the Chorley-Ribble dataset.	50
2.15	Chorley-Ribble log-risk function estimates utilising the ADAPT and FIXOS smoothing regimens, with MC tolerance contours (top row) and ASY tolerance contours (bottom row) superimposed at significance levels of 0.05 (solid line) and 0.1 (dashed line).	51
2.16	PBC cases (left) and population controls (right) for the PBC dataset.	52
2.17	PBC log-risk function estimates utilising the ADAPT and FIXOS smoothing regimens, with MC tolerance contours (top row) and ASY tolerance contours (bottom row) superimposed at significance levels of 0.05 (solid line) and 0.1 (dashed line).	53
3.1	Correlation functions from Table 3.1 plotted with $\alpha \in \{0.05, 0.1, 0.2, 0.4\}$; a legend is placed in the ‘Linear’ function plot.	67
3.2	Log-pair correlation functions g using as r the functional forms in Table 3.1, plotted with $\sigma^2 \in \{0.5, 1, 2, 4\}$ and a fixed scale $\alpha = 0.1$	68
3.3	A realisation of a LGCP with a Gaussian correlation structure (left) and a subsequently generated point pattern (right).	70
3.4	Nonparametric estimates of the transformed K - and pair correlation functions for the synthetic data, along with the parametric curves corresponding to the minimum contrast parameters $\hat{\xi}$	71
3.5	A rectangular grid approximating an irregular spatial region.	75
3.6	The extended rectangular grid in preparation for circulant embedding of cell-wise distances.	77
3.7	Comparing raw (left) and torus (right) Euclidian distance calculation between two distinct cell centroids on the extended lattice.	78
3.8	The extended lattice with irregular polygon wrapped on a torus in abstract space.	78
3.9	62 recorded spatial locations of childhood leukaemia and lymphoma in North Humberside, 1974–1986.	87
3.10	Fixed-bandwidth kernel intensity estimate of the global spatial variation of the Humberside data.	88
3.11	Minimum contrast estimation of the spatial correlation parameters for the Humberside data based on the K -function (left) and PCF (right).	89
3.12	Trace plots and diagnostics for the Humberside MALA run.	90
3.13	Posterior mean of the intensity surface $\hat{\lambda}(\mathbf{x})$ from the Humberside MALA run with original observations.	92
3.14	Spatial (left) and temporal (right) margins of the AEGISS data set.	93

3.15	Implemented deterministic global spatial trend $\hat{\zeta}_\psi$ (left) and global temporal trend $\hat{\eta}$ (right) for our AEGISS model.	94
3.16	Minimum contrast parameter estimation for the spatial dependence structure of the AEGISS data, via the time-averaged K -function (left) and PCF (right) for the selected Exponential correlation function.	94
3.17	Minimum contrast procedures for the spatial dependence structure of the AEGISS data, via the time-averaged K -function (left) and PCF (right), and ‘optimal’ curves based on the Gaussian and Stable correlation functions. . . .	95
3.18	Minimum contrast parameter estimation (Exponential correlation function) for the temporal dependence structure of the AEGISS data, based on comparing the parametric and nonparametric functions for the temporal covariance in the total number of cases per time point.	96
3.19	Trace plots of the realised, retained values of \hat{Z} for three distinct, randomly selected cells (identified simply as ‘blue’, ‘red’, and ‘green’) in the defined AEGISS space-time window for the MALA at $T = 1095$	98
3.20	Posterior mean intensity surfaces for each of $T - 4$ to $T - 1$ for the AEGISS analysis. Large circles are the observations on that day, small represent the locations of those from the previous day.	99
3.21	Posterior mean intensity surface for the most recent time T . Large circles are the observations on that day, small represent the locations of those from the previous day $T - 1$	100
3.22	95 th percentile empirical exceedance probabilities based on the realised exponentiated Gaussian fields for each of $T - 4$ to $T - 1$ for the AEGISS analysis.	101
3.23	95 th percentile empirical exceedance probabilities based on the realised exponentiated Gaussian fields for the most recent time point T	102
3.24	Examples of the kinds of point patterns produced under MCS1 to MCS9.	105
3.25	Absolute relative errors (ARE) for the mean parameter estimates produced from the simulations of MCS1 to MCS9, with respective labelling ‘K1’ to ‘K9’ for the K -function results, and ‘g1’ to ‘g9’ for the PCF results.	107
3.26	Deterministic spatial trends for Study 2.	108
3.27	Examples of the kinds of point patterns produced under DMCS1 to DMCS6, alongside uncorrelated realisations based on the global trends ‘A’, ‘B’ and ‘C’ alone.	110
3.28	Plotting the results of Study 2 for $n = 900$: DMCS1 to DMCS3. Variability of each mean is given as ± 2 s.e.; ‘ ’ for $\bar{\phi}$, ‘-’ for $\bar{\sigma}^2$. The pair minimising the RED is marked with a ‘×’. Points black→grey indicate increasing deterministic bandwidth. The location of the true parameter values is given with dashed lines.	112
3.29	Plotting the results of Study 2 for $n = 900$: DMCS4 to DMCS6. Variability of each mean is given as ± 2 s.e.; ‘ ’ for $\bar{\phi}$, ‘-’ for $\bar{\sigma}^2$. The pair minimising the RED is marked with a ‘×’. Points black→grey indicate increasing deterministic bandwidth. The location of the true parameter values is given with dashed lines.	113

3.30	Cutoff threshold value (95 th percentile) by variance σ_ψ^2 (solid, bold line). Values for STS1 (top), STS2 (middle) and STS3 (bottom), are given for the true variance (solid line), mean minimum contrast estimate by K (dashed line), and mean minimum contrast estimate by g (dotted line). Versions $T = 10; \eta = 100$ and $T = 100; \eta = 10$ are given in the left and right columns respectively.	120
3.31	True and ‘best’ predicted anomaly matrices (based on the surface with the highest Jaccard score, \mathcal{J}_{\max} , in each case) for both versions of STS1. Cells deemed anomalous as per the discussion are light, dark cells are those which are not.	123
3.32	True and ‘best’ predicted anomaly matrices (based on the surface with the highest Jaccard score, \mathcal{J}_{\max} , in each case) for both versions of STS2. Cells deemed anomalous as per the discussion are light, dark cells are those which are not.	124
3.33	True and ‘best’ predicted anomaly matrices (based on the surface with the highest Jaccard score, \mathcal{J}_{\max} , in each case) for both versions of STS3. Cells deemed anomalous as per the discussion are light, dark cells are those which are not.	125
4.1	Cases (red ‘o’) and controls (black ‘+’) for the PBC data.	137
4.2	Heatplots of the adaptive log-relative risk surfaces for the PBC data generated using sparr , displaying asymptotic and Monte-Carlo tolerance contours (left and right respectively), superimposed at significance levels of 0.01 (solid line) and 0.05 (dashed).	142
4.3	Two snapshots of the interactive 3D adaptive bandwidth PBC risk function, with superimposed asymptotic tolerance contours.	144
4.4	Plotted spatial coordinates and temporal margins (raw daily counts) of the AEGISS data from the corresponding R commands.	150
4.5	Interactive fixed-bandwidth density estimation of the AEGISS spatial data in lgcp	151
4.6	Showing the estimated global temporal trend $\eta(t)$	152
4.7	Interactive spatial parameter estimation in lgcp , based either on the PCF (top) or K -function (bottom).	154
4.8	Interactive temporal parameter estimation for θ^{-1} in lgcp , based on the the temporal covariance function.	155
4.9	Trace plots of the values of the retained Gaussian field for the three randomly selected torus cells (left) and an autocorrelation plot of the retained chain values for the third cell (right).	160
4.10	Posterior mean plots for the AEGISS data. Left: mean relative risk $\overline{\exp[Z_T]}$. Right: mean Poisson intensity $\zeta_\psi \overline{\eta \exp[Z_T]}$	162
4.11	Exceedance probabilities for the final time point for the AEGISS analysis in lgcp	164

5.1	Cases (●) and controls (+) for the Guichi schistosomiasis data, including study region (solid border line) and local waterways (dashed blue lines).	171
5.2	Initial risk function results for the schistosomiasis data, showing both fixed and adaptive versions with either MC or ASY upper-test tolerance contours plotted at a 5% level of significance.	172
5.3	Estimated cell-wise standard error (s.e.) surfaces based on the estimated fixed and adaptive risk functions, computed via both ASY and MC methods. . . .	174
5.4	Daily, weekly, and monthly H5N1 outbreak counts, showing temporal partitioning of the six epidemic waves. Image reproduced with kind permission from Z.J. Zhang.	177
5.5	Spatial adaptive kernel density estimates of each epidemic wave for the week-based H5N1 outbreak data set, also indicating significant clusters as per Kulldorff's space-time permutation scan statistic. Image reproduced with kind permission from Z.J. Zhang.	179
5.6	Map of Ireland, giving the primary infection location (×), silently infected premises (grey ○), and point of first detection (●), as well as the approximate boundary of the circular study region. Image ©2011 OIE publications; reproduced with permission.	182
5.7	Boxplots displaying the predicted counts of infected premises at eradication relative to the minimum count for each scenario and disease model. Image ©2011 OIE publications; reproduced with permission.	184
5.8	Averaged 5% asymptotic tolerance contours delineating heightened risk of infection of farms by FMD in Ireland, based on fixed-bandwidth kernel-smoothed risk surfaces, for each scenario and disease model. Images ©2011 OIE publications; reproduced with permission.	186
5.9	Spatial (left) and temporal (right) margins of the Japanese earthquake data. Spatial positioning of Japan's central land mass (grey shading) is visible in the former plot.	189
5.10	Assumed spatial and temporal global trends for the Japanese earthquake LGCP during the period 1926-2005.	191
5.11	Minimum contrast estimation of correlation parameters for the Japanese earthquake data over the period 1926-1965. Top: transformed K -function; center: transformed pair correlation function; bottom: temporal covariance function. Nonparametric versions are given a solid line, optimal parametric curves a dashed line.	192
5.12	Trace plot of the value of the Gaussian field at three random cell-locations: 'red', 'green' and 'blue' (left), and an autocorrelation plot for the 'blue' cell (right).	194
5.13	Estimating the spatial intensity for the Japanese earthquake data (2005). . .	195

List of Tables

2.1	Definitions of control densities g and risk functions r for the synthetic problem scenarios.	40
3.1	Some examples of stationary, isotropic correlation functions.	67
3.2	Study 1: MCS results, based on the K -function and PCF.	106
3.3	Definitions of global spatial trends used in Study 2.	108
3.4	Study 2: optimal pair of mean parameter estimates which have minimised the RED to the true values, with parenthesised standard errors, and corresponding deterministic bandwidth.	111
3.5	Study 3, Phase 1: Means and associated standard errors of the minimum contrast-estimated correlation parameters for both versions of each STS, based on the hypothetically observed data sets.	116
3.6	Study 3, Phase 2: Means and standard deviations of the Jaccard indices for both versions of each STS, based on MALA prediction of the final time point exceedances.	121

List of Publications

Accredited publications arising during the course of this research.

1. Taylor BM, **Davies TM**, Rowlingson BS, Diggle PJ (2012) **lgcp** - An R package for inference with spatial and spatiotemporal log-Gaussian Cox processes, *Journal of Statistical Software* [in press].
2. Zhang ZJ, Chen DM, Chen Y, **Davies TM**, Vaillancourt JP, Liu WB (2012) Risk signals of an influenza pandemic caused by highly pathogenic avian influenza subtype H5N1: Spatio-temporal perspectives, *The Veterinary Journal* **192** 3 417-421.
3. Zhang ZJ, **Davies TM**, Jiang QW (2012) Identification of high-risk regions for schistosomiasis in the Guichi region of China: an adaptive kernel density estimation-based approach, *Submitted for publication – Geospatial Health*.
4. **Davies TM**, Hazelton ML, Marshall JC (2011) **sparr**: Analyzing spatial relative risk using fixed and adaptive kernel density estimation in R, *Journal of Statistical Software* **39** 1 1-14.
5. Sanson RL, Harvey N, Garner MG, Stevenson MA, **Davies TM**, Hazelton ML, O'Connor J, Dubé C, Forde-Folle KN, Owen K (2011) Foot-and-mouth disease model verification and 'relative validation' through a formal model comparison, *Revue Scientifique et Technique-Office International des Epizooties* **30** 2 527-540.
6. **Davies TM**, Hazelton ML (2010) Adaptive kernel estimation of spatial relative risk, *Statistics in Medicine* **29** 23 2423-2437.
7. Hazelton ML, **Davies TM**, (2009) Inference based on kernel estimates of the relative risk function in geographical epidemiology, *Biometrical Journal* **51** 1 98-109 (**Pre-Ph.D.**).

List of Acronyms

Definitions of acronyms and abbreviations, and the section of their first occurrence in the thesis.

2D	Two-dimensional	1.2
3D	Three-dimensional	4.1.2
ADAPT	Adaptive-bandwidth relative risk function	2.5
AEGISS	Ascertainment and Enhancement of Gastrointestinal Infection Surveillance and Statistics	3.4.2
AMISE	Asymptotic mean integrated squared error	2.3.1
ARE	Absolute relative error	3.5.1
ASY	Asymptotic	2.4
BCV	Biased cross-validation	2.3.1
CRAN	Comprehensive R Archive Network	1.3
CSR	Complete spatial randomness	1.2.1
CV	Cross-validation	2.3.1
DFT	Discrete Fourier transform	3.3
DMCS	Deterministic Minimum Contrast Scenario	3.5.2
ETAS	Epidemic type aftershock sequence	5.4
EU	European Union	5.3.2
EW	Epidemic wave	5.2.2
FAO	Food and Agriculture Organisation of the United Nations	5.2.2
FFT	Fast-Fourier transform	3.3
FIXCV	Fixed-bandwidth relative risk function (CV bandwidth)	2.5
FIXOS	Fixed-bandwidth relative risk function (OS bandwidth)	2.5
FMD	Foot-and-mouth disease	5.3.1
GCMT	Global Centroid Moment Tensor	5.4
GPS	Global positioning system	5.1.2
iid	Independent and identically distributed	2.1.1
IQR	Interquartile range	2.3.1
ISE	Integrated squared error	2.3.1
KDE	Kernel density estimation	2.1
LGCP	Log-Gaussian Cox process	3.1
LOWESS	Locally weighted scatterplot smoothing	3.4.2
LSCV	Least-squares cross-validation	2.3.1
MALA	Metropolis-adjusted Langevin algorithm	3.3
MC	Monte-Carlo	2.4
MCMC	Markov-chain Monte Carlo	3.3.2
MCS	Minimum Contrast Scenario	3.5.1
MH	Metropolis-Hastings	3.3.2
MISE	Mean integrated squared error	2.3.1
NAADSM	North American Animal Disease Spread Model	5.3.1
netCDF	Network common data form	4.2.2
OIE	World Organisation for Animal Health	5.2.2
OS	Oversmoothing (maximal smoothing principle)	2.3.1
PBC	Primary biliary cirrhosis	2.6.2
PCF	Pair correlation function	1.2.3
PDF	Probability density function	2.1.1
RED	Relative Euclidian distance	3.5.2
SMR	Standardised mortality ratio	1.1
SNCP	Shot-noise Cox process	3.1.4
STS	SpatioTemporal Scenario	3.5.3
WISE	Weighted integrated squared error	2.5
WMISE	Weighted mean integrated squared error	2.3.1